Appendices

A Data collection and annotation

In this section, we discuss the process of collecting the text data that was used to train our machine learning classifiers presented in section 3 of the main manuscript. We provide several examples to illustrate how different classes of paragraphs are embedded in the judgment texts and discuss the inter-coder reliability of our data collection. As part of a wider effort to collect comprehensive data on the CJEU's case law (see Brekke et al. 2021), we hired four research assistants in March 2020, who were either pursuing or already completed a postgraduate degree in European Law at a Central and Northern European university. Our research assistants were tasked with reading the CJEU's judgments in preliminary reference procedures lodged with the Court between 1998 and 2011, extracting information from each paragraph in these judgments.[26] To facilitate the work of our research assistants, we programmed a web-application that allowed the research assistants to select individual judgments at a time for coding. Judgment texts were displayed by paragraphs and we provided our research assistants with drop-down selection menus listing our predefined labels of paragraph classes next to each paragraph.

Table 7 provides a representation of our web-application, illustrating how our research assistants saw the judgment texts for an excerpt of paragraphs from a randomly selected judgment of the CJEU, *Case C-152/10, Unomedical A/S v. Skatteministeriet.* The first column indicates the paragraph number (here, to save on space, starting with paragraph 23), the middle column displays the text of the paragraph, and the right column indicates the paragraph class chosen by the research assistant. The CJEU generally follows a template in writing its judgments in preliminary reference procedures. In the first paragraphs of the judgment, the Court summarizes the supranational and national legal contexts that are relevant for the case, and subsequently discusses the events that led to the main dispute in the referring national court. This discussion virtually always concludes with the observation that the national court decided to stay its proceedings and refer the case for a preliminary ruling to the CJEU (see paragraph 23 in Table 7 for an illustrative example).

[26] In addition to identifying the correct paragraph class, for each paragraph our research assistant were tasked with identifying the CJEU's references to European and international law as well as references to observations that had been submitted to the Court by Member States, EU institutions and third parties.

Table 7: Extract of paragraph texts and classes from *Case C-152/10, Unomedical A/S v. Skatteministeriet*

Number	Paragraph Text	Paragraph Class
23	Taking the view that the classification of the bags at issue in the main proceedings depends on the interpretation of the notions of 'parts' and 'accessories' in Chapter 90 of the CN, the Højesteret decided to stay the proceedings and to refer the following questions to the Court for a preliminary ruling: 1. Is a dialysis bag, manufactured from plastic, which is specially designed for, and can only be used with, a dialyser to be classified under: – Chapter 90 [...] or – Chapter 39 [...] as plastics or articles thereof? 2. Is a urine drainage bag, manufactured from plastic, which is specially designed for, and therefore can only be, and in fact is, used exclusively in connection with, a catheter, to be classified under: – Chapter 90 [...] or – Chapter 39 [...] as plastics or articles thereof?'	`residual`
24	By its two questions, which it is appropriate to consider together, the national court asks the Court, in essence, whether, for the period from 1 May 2001 to 31 December 2003, plastic drainage bags were to be regarded as 'parts' and/or 'accessories' for a catheter or a dialyser and, therefore, classified under heading 9018 of the CN, or whether they were to be classified under heading 3926 of the CN as 'articles of plastics'.	`question_start`
25	At the outset, it should be borne in mind that, according to the Court's settled case-law, in the interests of legal certainty and for ease of verification, the decisive criterion for the classification of goods for customs purposes is in general to be sought in their objective characteristics and properties as defined in the wording of the relevant heading of the CN and the notes to the sections or chapters [...].	`residual`
(...)	(...)	(...)
42	Furthermore, it must be stated that, contrary to the argument advanced by Unomedical, those interpretative opinions, which have not led to the adoption of a regulation, can validly be used in legal relationships which arose and were established before those opinions were adopted.	`residual`
43	Consequently, the answer to the questions referred is that the CN must be interpreted as meaning that a dialysis drainage bag, manufactured from plastic, which is specially designed for, and can be used only with, a dialyser (artificial kidney), had, between May 2001 and December 2003, to be classified under subheading 3926 90 99 of the CN as 'plastics and articles thereof' and that a urine drainage bag, manufactured from plastic, which is specially designed for, and therefore can be used only in connection with, a catheter had, during the same period, to be classified under subheading 3926 90 99 of the CN as 'plastics and articles thereof'.	`question_stop`

The table provides a representation of the web-application used by our research assistants to code paragraph classes, and provides illustrative examples of how the paragraph classes `question_start` and `question_stop` are embedded in the judgment text.

The CJEU then summarizes and occasionally combines or rephrases the national court's question(s). We instructed our research assistants to identify at which paragraph the CJEU transitioned from summarizing the proceedings in the national court to providing its answer to the legal issue at hand, and assign the class `question_start` to this paragraph. Paragraph 24 in Table 7 provides an illustrative example of this paragraph class. In the preceding paragraph, the CJEU noted that the referring Danish court had stayed its proceedings and referred two questions concerning the appropriate classification of medical plastic drainage bags to the CJEU. In paragraph 24, the CJEU then summarizes and combines the Danish court's questions before moving on to elaborate on these questions in paragraph 25 and subsequent paragraphs (not displayed in Table 7).

Once our research assistants had classified a paragraph as `question_start`, they were instructed to continue to read through the judgment text, until they arrived at a paragraph that provided the CJEU's conclusive answer to the national court's question(s). Our research assistants were instructed to then classify this paragraph as `question_stop`. Paragraph 43 in Table 7 provides an illustrative example of this paragraph class, providing a conclusive answer to the Danish court's questions that had been summarized in paragraph 24. The CJEU concluded that both dialysis and urine drainage bags had to be classified as 'plastics and articles thereof'. Based on these classifications, paragraphs 24 to 43 belong to the same distinct legal issue considered by the Court in *Case C-152/10, Unomedical A/S v. Skatteministeriet.*

As discussed in the main manuscript, a key contribution of our approach is the ability to distinguish between several legal issues within a single judgment and to identify the paragraphs that address these distinct issues, respectively. In the following, we provide an illustrative example of the CJEU's consideration of multiple legal issues within a single judgment in *Case C-127/08 Blaise Baheten Metock and Others v Minister for Justice, Equality and Law Reform.* Table 8 shows a selection of paragraphs from the judgment. The CJEU had been asked by the High Court of Ireland to provide answers to three distinct questions. Paragraph 80 provides the Court's answer to the High Court's first question, concluding that Member State legislation that specifies certain residence requirements for spouses of Union citizens is precluded by Directive 2004/38.[27]

[27]The corresponding paragraph of the class `question_start` reads "By its first question the referring court asks whether Directive 2004/38 precludes legislation of a Member State which requires a national of a non-member country who is the spouse of a Union citizen residing in that Member State but not possessing its nationality to have previously been lawfully resident in another Member State before arriving in the host Member State, in order to benefit from the provisions of that directive."

Table 8: Extract of paragraph texts and classes from *Case C-127/08 Blaise Baheten Metock and Others v Minister for Justice, Equality and Law Reform*

Number	Paragraph Text	Paragraph Class
80	The answer to the first question must therefore be that Directive 2004/38 precludes legislation of a Member State which requires a national of a non-member country who is the spouse of a Union citizen residing in that Member State but not possessing its nationality to have previously been lawfully resident in another Member State before arriving in the host Member State, in order to benefit from the provisions of that directive.	`question_stop`
81	By its second question the referring court asks essentially whether the spouse of a Union citizen who has exercised his right of freedom of movement by becoming established in a Member State whose nationality he does not possess accompanies or joins that citizen within the meaning of Article 3(1) of Directive 2004/38, and consequently benefits from the provisions of that directive, irrespective of when and where the marriage took place and of the circumstances in which he entered the host Member State.	`question_start`
82	It should be noted at the outset that, as may be seen from recitals 1, 4 and 11 in the preamble, Directive 2004/38 aims to facilitate the exercise of the primary and individual right to move and reside freely within the territory of the Member States that is conferred directly on Union citizens by the Treaty.	`residual`
(...)	(...)	(...)
98	Third, neither Article 3(1) nor any other provision of Directive 2004/38 contains requirements as to the place where the marriage of the Union citizen and the national of a non-member country is solemnised.	`residual`
99	The answer to the second question must therefore be that Article 3(1) of Directive 2004/38 must be interpreted as meaning that a national of a non-member country who is the spouse of a Union citizen residing in a Member State whose nationality he does not possess and who accompanies or joins that Union citizen benefits from the provisions of that directive, irrespective of when and where their marriage took place and of how the national of a non-member country entered the host Member State.	`question_stop`
100	In view of the answer to the second question, there is no need to answer the third question.	`question_noanswer`

The table provides a representation of the web-application used by our research assistants to code paragraph classes, and provides illustrative examples of paragraphs marking the transition between two legal issues within a single judgment, as well as paragraphs indicating that the CJEU provided no answer to the national court's question, the class `question_noanswer`.

Following its answer to the first question in paragraph 80, classified as `question_stop`, the Court then moves on to summarize the Irish High Court's second question in paragraph 81, to be classified as `question_start`, marking the transition from the first legal issue to the second legal issue in the judgment text. The second legal issue then concludes with the CJEU's answer to the High Court's second question in paragraph 99, classified as `question_stop`. Based on these classifications we can then distinguish between two sets of paragraphs within the judgment text that address distinct issues, paragraphs 48 (not shown here) to 80 for the first legal issue, and paragraphs 81 to 99 for the second legal issue.

Finally, as noted in the main manuscript, the CJEU occasionally concludes that a question referred by a national court does not require an answer, typically because the CJEU argues that a previous answer had already sufficiently addressed the national court's question. Paragraph 100 in Table 8 provides an illustrative example of the paragraph class `question_noanswer` and its typical embedding in the judgment text, immediately following up on the paragraph concluding the previous legal issue.

For judgment texts similar to the illustrative examples discussed above, we can reasonably expect that trained research assistants familiar with European law are capable of correctly distinguishing between the different paragraph classes and code individual paragraphs accordingly. However, at the start of our research project we were unsure whether we would consistently find the kinds of patterns identified in the illustrative examples, recognizing that the absence of such patterns would make it more likely that different research assistants would classify paragraphs differently within the same judgment text. We therefore designed two rounds of reliability checks, allowing for double-blinded coding of judgment texts by pairs of research assistants. In the first round of reliability checks, we randomly assigned 100 judgment texts to pairs of research assistants, with assignments overlapping across research assistants (i.e. different subsets of the judgments assigned to research assistant A were also assigned to research assistants B, C, and D, respectively). After the first round of reliability checks, we evaluated the inter-coder reliability for paragraph classifications, discussed and resolved instances of disagreements with our assistants, and then repeated the process for another set of 50 judgment texts randomly assigned to pairs of research assistants.

Following the double-blinded coding, we were able to compare the paragraph classifications of pairs of research assistants and calculate the proportion of matching classifications (i.e. two research

Paragraph class	First round	Second round	Total
question_start	0.84 (157 of 186)	0.93 (80 of 86)	0.87 (237 of 272)
question_stop	0.86 (167 of 194)	0.95 (79 of 83)	0.89 (246 of 277)
question_noanswer	1.00 (16 of 16)	0.88 (8 of 9)	0.96 (24 of 25)

Table 9: Scores indicate the proportions of identical coding decisions per paragraph class (frequencies are reported in parentheses).

assistants independently chose the same class for the same paragraph). The 100 judgments selected for the first round of reliability checks comprised a total of 4,420 paragraphs. We found matching classifications in 98 percent of these paragraphs. The 50 judgments selected for the second round of reliability checks comprised a total of 2,189 paragraphs and we found matching classifications in 99 percent of these paragraphs.

As noted in the main manuscript, most paragraphs in judgment texts are classified as `residual` and the frequent occurrence of one particular paragraph class (the one we are also least interested in) is likely to inflate the proportion of matching classifications. We therefore calculated the proportions of matching classifications for our three paragraph classes of interest, `question_start`, `question_stop` and `question_noanswer`, respectively. Table 9 reports the results. We can see that the proportion of identical coding decisions was relatively high across all paragraph classes of interest, with the lowest proportion coming in at 0.84 for the paragraph class `question_start`. These results further improved for the second round of reliability checks, after our research assistants had undergone additional training (save for the paragraph class `question_noanswer`). Overall, we acknowledge that our coding instructions do not fully eliminate the possibility of differences in our research assistants' coding of paragraph classes, and that these differences feed into the performance of our machine learning classifiers. However, based on the results of our reliability checks, particularly after research assistants had undergone additional training, we expect that the extent of these differences is small.

B Model specification and tuning

In the following, we discuss the hyper-parameter space for two of our machine learning classifiers, the feedforward neural network and the Random Forest model, and provide further information

Table 10: Network structure

Layer (Type)	*Output shape*	*# parameters*
Input layer (Dense) with $\ell 2$-regularization $\lambda = 0.0001$	Batch size = 64, Nodes = 64	651,200
Dropout layer with dropout rate = 0.3	Batch size = 64, Nodes = 64	–
Hidden layer (Dense) with $\ell 2$-regularization $\lambda = 0.0001$	Batch size = 64, Nodes = 64	4,160
Dropout layer with dropout rate = 0.3	Batch size = 64, Nodes = 64	–
Output layer (Dense)	Batch size = 64, Nodes = 4	260

Note: Structure of our feedforward neural network with a single hidden layer. Hyper-parameters were tuned using random sampling.

on the tuning processes we employed to identify the optimal hyper-parameters for these classifiers. We then compare the performance of our bag-of-words classifiers to two networks that incorporate the sequence of features, a one-dimensional deep convolutional neural network (CNN) and a long short-term memory (LSTM) network.

B.1 Feedforward neural network

In the following, we provide information on the architecture and hyper-parameters of our feedforward neural network. We also discuss the tuning process we used to identify the optimal combination of hyper-parameter values and the stability of the network's estimates.

We programmed a feedforward artificial neural network with a single hidden layer, using the **keras** and **tensorflow** packages for R. To avoid overfitting, we implemented $\ell 2$-regularization in both the input and hidden layer. Further, we added two dropout layers, one after the input layer and another after the hidden layer. Table 10 illustrates the architecture of our neural network. The input layer uses a rectified linear unit activation function (ReLU) while the output layer uses a softmax function, returning a probability distribution over the four paragraph classes for each paragraph. Since we are solving a multi-class classification problem we are using a categorical cross entropy loss function and choose a Root Mean Square Propagation (RMSProp) optimizer.

To identify the optimal values for the network's hyper-parameter space we randomly sampled different combinations of hyper-parameter values, comparing the classification performances of the network for different values for the dropout rate (with rates varying between 0.2, 0.3 and 0.4), the

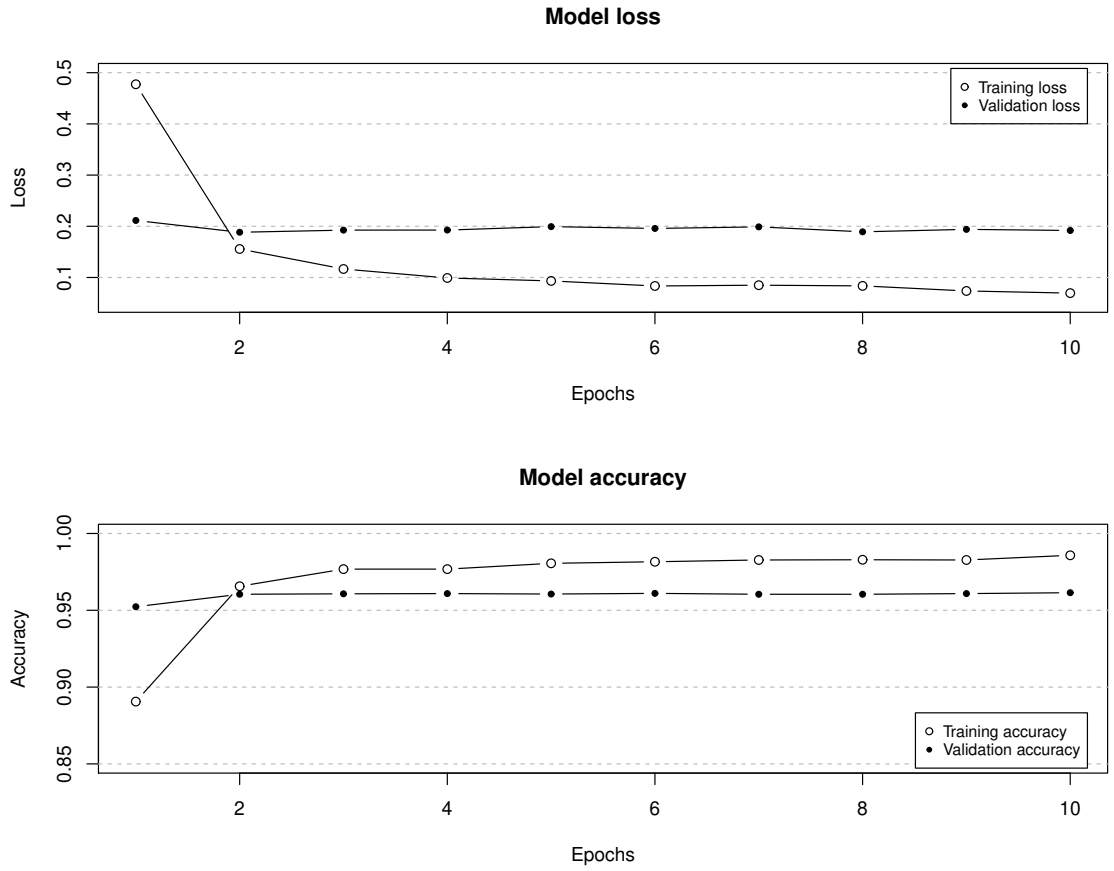

Figure 9: Loss and accuracy of our fitted neural network for the training and test set across ten epochs.

number of nodes in the input and hidden layers (with the number of nodes varying between 64, 128 and 256), the $\ell 2$-regularization parameter λ (varying between 0.0001, 0.001 and 0.01), as well as the learning rate for the RMSProp optimizer (with rates varying between 0.0001, 0.001 and 0.01). We sampled 20 percent of the possible combinations and found that a network with dropout rates in both dropout layers set at 0.3, 64 neurons in the input and hidden layer, $\lambda = 0.0001$, and the RMSProp optimizer's learning rate set at 0.001 provides the highest classification accuracy for our test set.

We fit the neural network with the identified optimal hyper-parameters for ten epochs, which takes roughly two minutes on our CPU. Figure 9 plots the network's loss and accuracy for the training and validation set across the ten epochs. We can see that changes to the validation loss and accuracy are minimal after the second epoch, and that the network settles on a classification

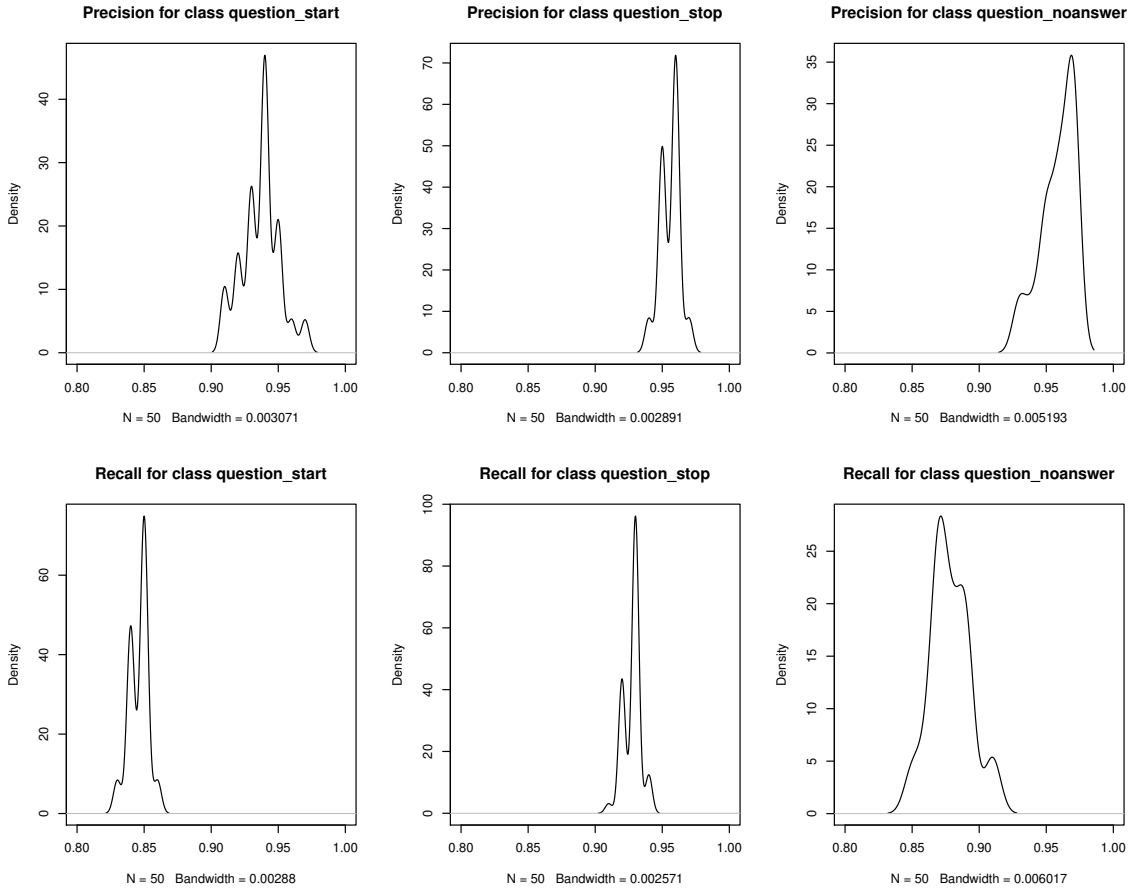

Figure 10: Distributions of precision and recall rates for the paragraph classes `question_start`, `question_stop` and `question_noanswer` across 50 iterated runs of our feedforward neural network.

accuracy around roughly 96 percent after the second epoch. Given the stochastic nature of the learning algorithm, the neural network's classification performance varies from run to run. To identify whether the network's performance remains stable across different runs, we estimated the neural network for 50 iterations, each time recording the precision and recall rates for the paragraph classes. In Figure 10 we plot the distributions of precision and recall rates for the three paragraph classes of interest, `question_start`, `question_stop` and `question_noanswer`. Figure 10 shows that the estimates are relatively stable, with only recall rates for the class `question_start` regularly dipping below 0.85, yet consistently staying well above 0.80.

All files and code to replicate the analyses presented here and in the main manuscript are made available via the supplementary materials.

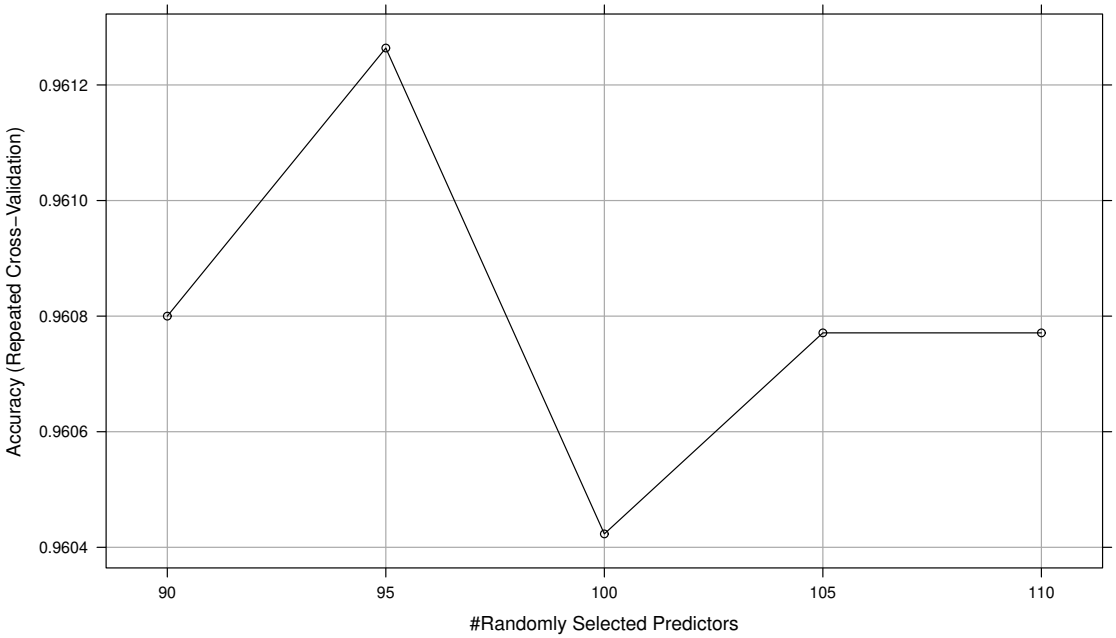

Figure 11: Comparison of classification accuracy for Random Forest models with varying numbers of features randomly sampled at each split (i.e. the models' *mtry* parameter). Results are based on five-fold repeated cross-validation.

B.2 Random Forest model

Our discussion of the hyper-parameter tuning of our Random Forest model centers on two parameters: (1) the number of features in our document-feature matrix available for splitting at each tree node (in the following referred to as *mtry* in line with the corresponding parameter name in the **randomForest** package), and (2) the number of trees to grow (in the following referred to as *ntree*). Larger values for the *ntree* parameter improve model stability. However, in light of the dimensions of our document-feature training matrix (a $6,899$ paragraph $\times$ 10,175 three-grams matrix) and the hardware available to us, we opted for $ntree = 25$ as larger values on the parameter meant that the Random Forest model could not be estimated due to insufficient memory on our machine.

To choose the optimal value for the *mtry* parameter, we employed a grid-search with five-fold cross-validation and five repetitions, implemented through the **caret** package for R. The default value for the *mtry* parameter is the square root of the number of features in our training matrix, which is roughly 100 in our case. For our tuning grid, we considered values clustered around this default value: 90, 95, 100, 105 and 110. Figure 11 plots the results of the grid search, identifying

mtry = 95 as the optimal value. Based on these results, we specify a Random Forest with *ntree* = 25 and *mtry* = 95. The minimum size of terminal nodes is left at its default value of 1 and we don't impose any limits on the maximum number of terminal nodes trees of the Random Forest model. It takes roughly 35 minutes to run the model on our standard CPU machine.

B.3 CNN and LSTM networks

In the following, we discuss the performance of two additional classifiers that follow a different approach than the bags-of-words classifiers discussed above. We programmed a one-dimensional convolutional neural network (CNN) and a long short-term memory (LSTM) network to solve our classification task. These networks incorporate the sequence of features in their learning algorithms. Our bags-of-words classifiers take sparse document-feature matrices as their input, where each document (here, each paragraph) is represented by a row vector of integers, indicating the number of times the respective feature (with one feature per column in the matrix) occurs in the document. The CNN and LSTM network, on the other hand, rely on word embeddings, where each word in the text is represented by a real valued vector in a high-dimensional space—put simply, words with a similar meaning have similar representations in the vector space (i.e. their vector representations point into similar directions), and the networks can learn these representations during training.

We start by splitting our data for each paragraph class in half to create our training and test set. We then encode each paragraph into a sequence of integers, with each integer mapping to one specific feature in the vocabulary of our paragraphs. To do so, we use the text_tokenizer and text_to_sequences functions of the **keras** package for R, and limit the vocabulary to the 10,000 most frequently occurring words in the paragraphs while filtering out punctuation and lowercasing all words. The implementation of the CNN and LSTM network through the **keras** package requires that the integer sequences for all paragraphs have the same length. We identify the maximum number of words in the paragraph classes of interest to us (here, a paragraph of the class `question_stop` with 1,099 words), zero-pad all shorter paragraphs to the length of 1,099, while cutting all words after the 1,099th word for all longer paragraphs.

For our CNN, we start with an embedding layer, specifying a 100-dimensional vector space for the features in paragraphs. We follow the embedding layer with a 1D convolution layer with 64 output filters and kernel size 8, and choose a ReLU activiation function. We then include a pooling

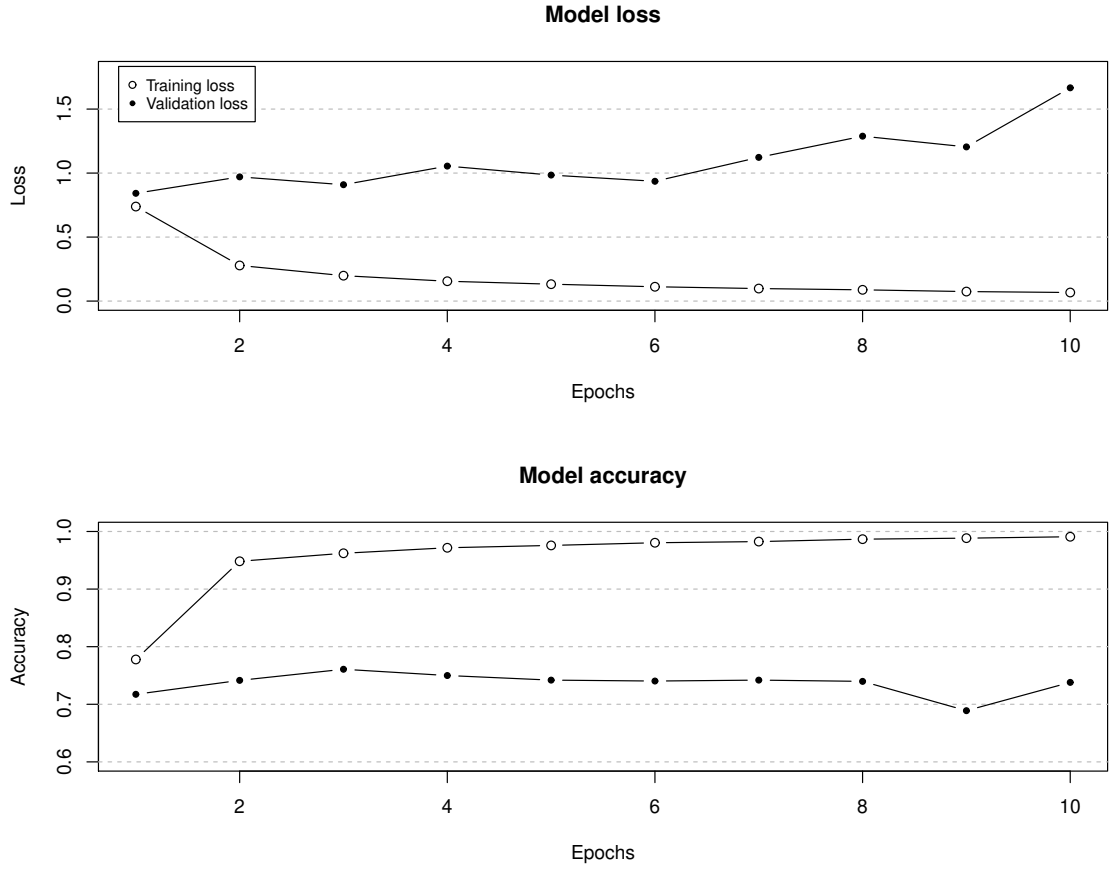

Figure 12: Loss and accuracy of our fitted convolutional neural network for the training and test set across ten epochs.

layer with two pooling windows, a layer to flatten the input, and standard hidden layer with 64 filters, again using a ReLU activation. As for our feedforward neural network, the output layer uses a softmax function, returning a probability distribution over the four paragraph classes for each paragraph. We are using a categorical cross entropy loss function and choose a RMSProp optimizer as our optimization algorithm. We fit the CNN for 10 epochs with a batch size of 64 which takes roughly 5 minutes on our CPU, monitoring loss and accuracy for the training and test set, plotted in Figure 12. We can see in Figure 12 that the CNN struggles to accurately classify paragraphs in the test, with the maximum test set accuracy, roughly 0.78, falling well below the performance of the artificial feedforward neural network and the Random Forest model. We tried several other configuration for the CNN not reported here yet none of these configurations came close to achieving a similar performance as reported for our bag-of-words approaches.

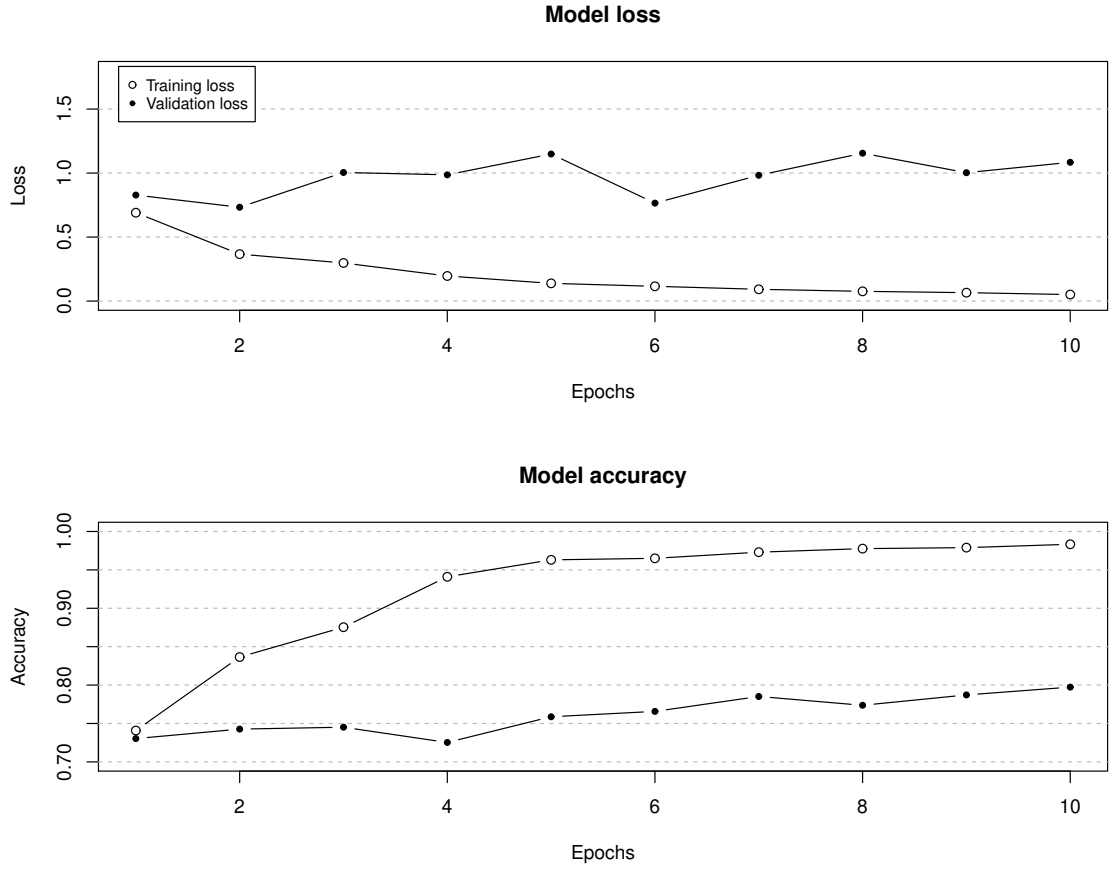

Figure 13: Loss and accuracy of our fitted LSTM network for the training and test set across ten epochs.

We find a similar pattern for our LSTM network. For the configuration reported here, we followed the embedding layer with a dropout layer, with the dropout rate set at 0.25. We then included a 1D convolution layer with 64 filters kernel size 8 and a ReLU activation, followed by a pooling layer with two pooling windows. We then include a long short-term memory layer with 64 units, followed by the output layer with 4 units and a softmax activation. As before, we choose a categorical cross entropy loss function and a RMSProp optimizer as our optimization algorithm. We fit the LSTM network for 10 epochs which takes roughly 12 minutes on our machine and plot loss and accuracy for the training and test set in Figure 13. While test set accuracy gradually improves up to 0.82 in the 10th epoch, the LSTM network (as well as other configurations we tried, not reported here) still falls short of the performance of the bag-of-words approaches.

The CNN and LSTM network do not outperform classifiers using bags-of-words for the data

we are working with—the CJEU's judgments in preliminary references issued between 1998 and 2011—yet we want to stress that this may not be the case for other applications. As discussed in the main manuscript, the paragraph classes we are interested in are characterized by relatively short sequences of words often at the beginning of the paragraph, e.g. "the referring court is asking", which we can easily capture using n-grams, while the sequence of features in the remaining paragraph text matters less. Classifiers relying on bags-of-words appear well-suited to solve classification tasks with such patterns. But the patterns that characterize paragraphs beginning or concluding the discussion of an issue are likely to look different for other courts, and classifiers incorporating the information on the sequence of words may be the better choice for these applications.

C Topic modelling

As described in Section 4.1, we train a LDA topic model on a subset of judgments. For each judgment, both the text of entire judgments as documents and the text of the issues as document are included to eliminate any bias for either approach. We set the number of topics $k = 10$. In Figure 14 we plot word clouds that highlight the distribution of tokens across the ten different topics, and allow us to evaluate the face validity of the topics identified by the model. For reasons of legibility, the figure is limited to the 300 most frequently occurring tokens.

From our understanding of EU law, the tokens associated with each topic are connected to each other and to known major areas of CJEU free movement of goods jurisprudence. Some topics are straightforward to interpret. For example, it is clear from the associated words that certain topics relate to particular types of products. For example, Topic 1 concerns health care, pharmaceutical products, and products harmful to public health care, incl. the marketing of such products, Topic 4 products related to broadcasting, and Topic 5 agricultural and other food products. Other, non-product related topics are also straightforward to interpret. Topic 6 concerns the protected designation of products from certain regions and related issues, while Topic 2 concerns direct taxation of certain products, as well as discriminatory and inequitable taxation on goods from other Member States. Topic 3 concerns the classification of products based on their functions and characteristics, Topic 8 custom duties, and Topic 10 shipping and other forms of transport. The remaining two topics are somewhat more elusive. Overall, however, the topics identified by the

Figure 14: The frequency of the 300 most frequent tokens across topics. Size represents frequency and colors different topics.

LDA topic model appear to capture various, central aspects of EU law on free movement of goods that we would expect to find in this corpus, increasing our confidence in the validity of the model.

Each topic model trained is somewhat different, even if identical train and test data is used for training, and the manual confirmation offered here applies to one and only one such possible topic model.

This also means that the improvement offered by issue-splitting on topic modelling will vary dependent on the model. To ensure that the improvement offered by our approach are robust, we trained 100 topic models (using the same train and test data) and applied each model to a selection judgments not included in the training data (test data). For each, we captured the maximum

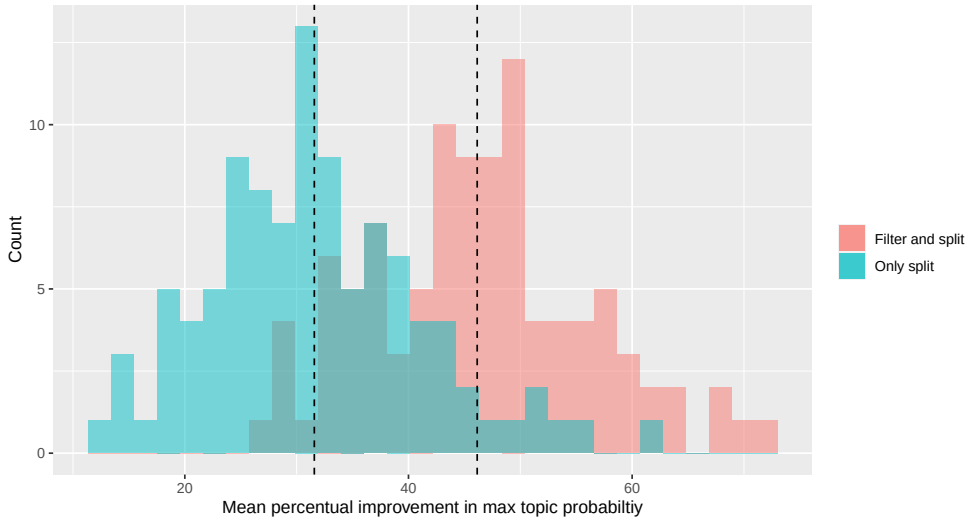

Figure 15: The figure shows percentage increase in mean maximum topic probability for issues compared to entire judgments ("split and filter") and all issues together ("only split"). Each observation is the improvement under one of 100 generated topic models. Vertical lines show averages for the two groups.

topic probability for each issue and calculated by what factor this value was greater than the same model applied to (i) the completed judgment text to measure the improvement from both filtering and issue-splitting and (ii) the combined text of the issues to measure the improvement from only issue-splitting.

The results of these tests are displayed in the figure below. There is variance in the improvement offered. However, all tests showed a benefit and filtering and issue-splitting was consistently better than issue-splitting. Average across all 100 tests, the improvement offered by issue-splitting was 31.6% and filtering and issue-splitting was 46.1%.

D Network analysis

In this section we describe in greater detail the network or graph analysis elements of the study.

We began by identifying from the text of the studied judgments (source) all references to other, previous CJEU judgments (target), including both judgments that are part of the studied sample and those which are not. We recorded in which paragraph of the referring judgment the reference is found and, where available, the referred to paragraph of the target judgment. The sample contains 2,476 such references. We treat those references as edges in a time-directed network.

Table 11: Three edge versions

	Source	Target
Paragraph-level edge	ECLI:EU:C:2009:521–98	ECLI:EU:C:2003:295-37
Issue-level edge	ECLI:EU:C:2009:521:I3	ECLI:EU:C:2003:295:I1
Judgment-level edge	ECLI:EU:C:2009:521	ECLI:EU:C:2003:295

Take for example the following excerpt from in Case C-478/07, *Budějovický Budvar, národní podnik v. Rudolf Ammersin GmbH*, ECLI:EU:C:2009:521, para. 98:

> It follows that, since the bilateral instruments at issue now concern two Member States, their provisions cannot apply in the relations between those States if they are found to be contrary to the rules of the Treaty, in particular the rules on the free movement of goods (see, to that effect, Case C-469/00 *Ravil* [2003] ECR I-5053, paragraph 37 and the case-law cited).

As presented in Table 11, the references from *Budvar* to *Ravil* can be represented as an edge in three different ways that are all accurate but differ in terms of specificity. The paragraph-level edge is the most specific. By discarding the paragraph information we can easily create judgment-level edges. However, we can also use the paragraph information to identify the source and target issue respectively which, in turn, allows us to create issue-level edges of intermediate specificity. Thus, using the same information we are able to create case law citation networks of varying degree of specificity: a paragraph-level network (not part of the study), an issue-level network, and a judgment-level network.[28]

We then calculate centrality for all vertices in both the issue-level and judgment-level network. We seek to capture the persuasiveness of the Court's reasoning and, more specifically, how well embedded it is in existing jurisprudence. There are a few centrality measurements worth considering for this purpose, each with certain advantages and disadvantages.

The most simple and straight-forward measurement is *outdegree* which is equal to the number of out-going references. The main advantages with outdegree is that it does not change[29] and

[28]This means that only judgments and issues that contain references are included in the networks and are capable of receiving centrality scores.

[29]The number of references in a judgment are and always will remain the same regardless of how case law develops.

that it is calculated locally, i.e. it is independent of the rest of the network and therefore also of sampling. Them main drawback with outdegree is that it (in its basic form) attaches equal weight to all edges. In the context of this study it means that it only captures the quantity of references without regard to quality, which is non-ideal.

For this reason, outdegree is frequently replaced or supplemented by *hub score*. Hub score, which is one side of the HITS-algorithm, was developed for the purpose of identifying web pages that link to good authorities (Kleinberg 1999). Like outdegree, a vertex's hub score is based on its outward edges, but instead of only reflecting how many other vertices a vector is connected to hub score also reflects how many other vertices point to those target vertices. Thus, hub score incorporates a qualitative element not present in outdegree. However, hub score tends to perform less well on small data sets and is sensitive to small changes.

We here consider if and to what extent the results differ if instead of outdegree we use hub score or hub rank, a variant calculated on the basis of target vertices page rank rather than authority (see Derlén and Lindholm 2017).

E Additional results

In the following, we present additional results from our analyses, considering alternative measures for network centrality for our outcome variable *Outdegree* discussed in the previous section, *Hub Score* and *Hub Rank*.

Both *Hub Score* and *Hub Rank* are non-negative continuous variables, with values concentrated in the left tails of their distributions (this is particularly true for *Hub Score*). Figure 16 plots of both variables for the judgment- and issue-level. We can see that there is hardly any variation on the variable *Hub Score* for both levels (with most values concentrated at or only marginally above zero), underlining concerns that calculating hub scores to measure network centrality is infeasible for relatively small datasets like ours. In light of these distribution, we have reason to expect that the estimated hub scores do not reflect a valid measure of network centrality for our data and we therefore proceed by estimating our models for robustness checks only for the outcome variable *Hub Rank*.

Given the variable *Hub Rank* is continuous with a right-skewed distribution, we opt for Gen-

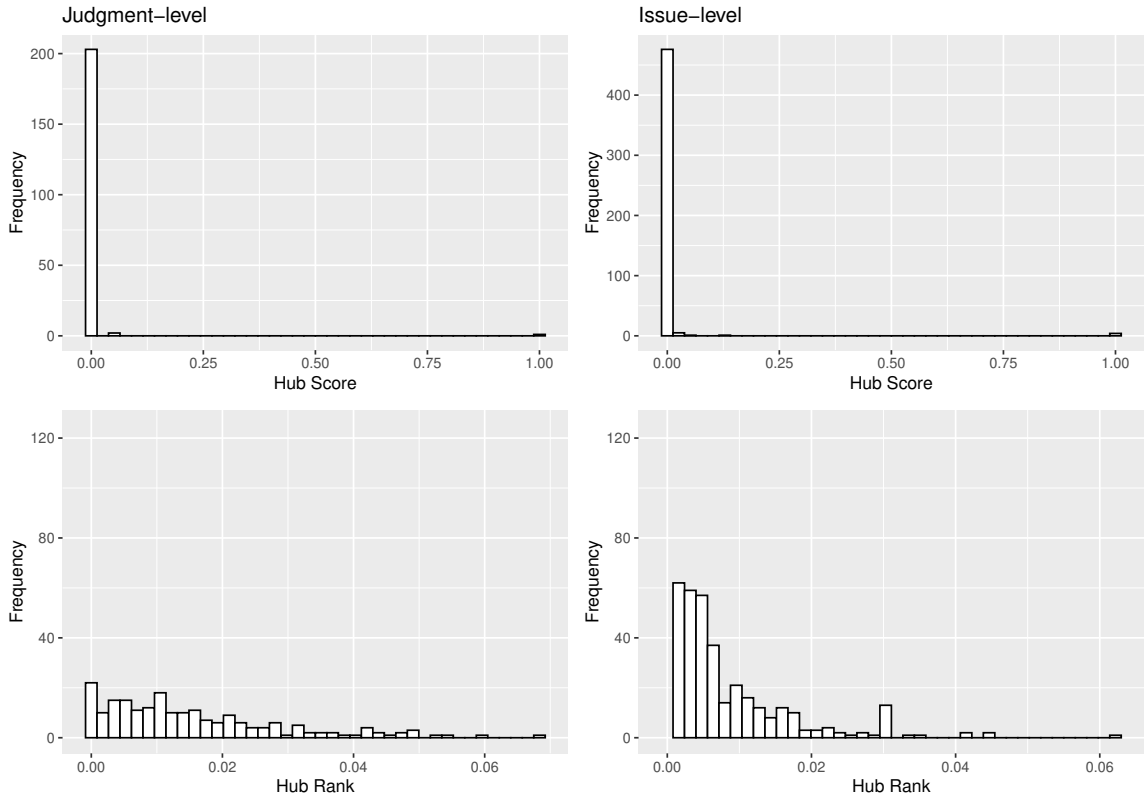

Figure 16: Distributions for outcome variables *Hub Score* and *Hub Rank* at the judgment-level ($N = 206$) and issue-level ($N = 487$).

eralized Linear Models with a Gamma distribution and a log link function for our regressions. To account for the zero-values on many of our observations at both the judgment and issue-level, we opt for a linear transformations of the outcome variable, adding a small value of 0.01 to each observation.

Coefficient estimates for the judgment and multi-level regression models are displayed in Figure 17, along with their 95% HPDs. Overall, we find that coefficient estimates for the models including *Hub Rank* as outcome variable are by and large similar to the results discussed in Section 5.3 of the main manuscript. Notably, coefficients for the category *In conflict* of the variable *MS Conflict* remain positive and distinguishable from zero. In addition, similar to the results reported in the main manuscript, coefficient estimates for the category *In conflict* of the variable *AG Conflict* are positive and distinguishable from zero. These results suggest that the CJEU is more likely to cite existing case law with higher precedential authority when its position runs counter to the positions expressed by Member States and the Advocate General, respectively.

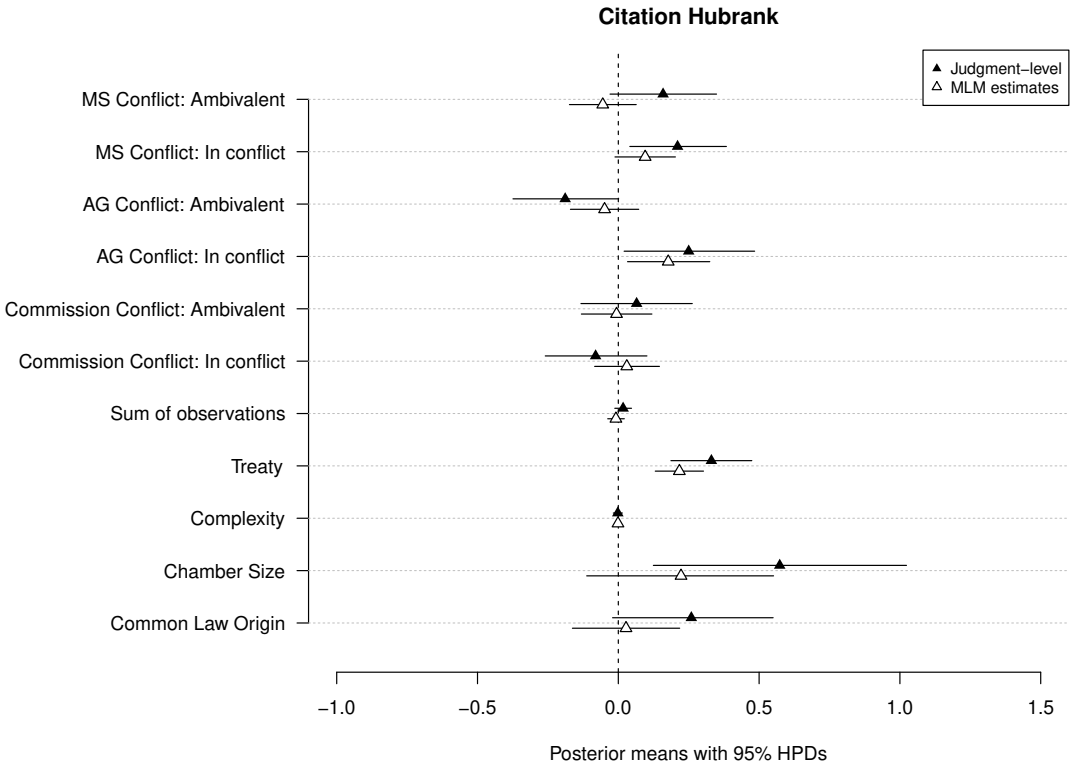

Figure 17: Posterior means with 95% HPD intervals of regression coefficients, displayed for the judgment-level ($N = 206$) and multi-level analyses ($N = 487$). All regression analyses include year fixed-effects (not shown here).

Again, similar to the results discussed in section 5.3 of the main manuscript, the main differences for the judgment-level and multi-level regression can be observed for the coefficients *MS Conflict: Ambivalent* and *AG Conflict: Ambivalent*. The judgment level regression would suggest that the CJEU is *more* likely to reference case law with higher precedential authority when the Court and Member States' position concerning a further restriction of national autonomy is ambivalent, and *less* likely to do so when the Court and Advocate General's positions are ambivalent. These effects, however, disappear once we consider the actual issues the CJEU addressed in its judgments and avoid aggregating the actors' positions to the judgment-level, mirroring results from section 5.3.